Discussion of *The Missing Value of Data*

Bhutani, Ordoñez, and Veldkamp (May 2026)

Dimitris Papanikolaou

NBER Macro Annual, Spring 2026

1 Introduction

A growing literature has established that data is a valuable economic asset (Jones and Tonetti, 2020; Farboodi and Veldkamp, 2021; Abis and Veldkamp, 2024). The current paper builds on this agenda and has an ambitious goal: estimate the omitted contribution of data to aggregate GDP. It starts with the novel observation that, if data is used as a method of payment, national accounting misses the value of data **twice:** it misses the cost of creating the data, and it misses the fact that data is bartered for goods. The first channel is not unique to data, as most intangible investment suffers from the same omission. The second channel is genuinely novel. The paper proposes a simple measurement methodology to estimate the stock of data based on the quality of firms' revenue forecasts. Then, using a calibrated structural model, it maps the inferred stock of data into missing value-added from the national accounts.

2 Comments

My comments center on the measurement methodology. I organize them around three questions. The first is whether managerial revenue forecasts are a clean proxy for the stock of data—whether firms use data in the narrow way the model assumes, and whether reported guidance is the rational forecast the measurement presumes or instead embeds a systematic, predictable bias. The second is identification: the calibration treats the relation between gross profits and forecast errors as causal, yet the same slope is consistent with omitted management quality—a confounder distinct from the reverse-causality feedback the model explicitly addresses. The third concerns the model's distributional assumptions: I ask how

the measured stock of data—and its behavior over the business cycle—would change once we relax the assumptions of a symmetric loss function and normal distribution.

2.1 Measurement

The paper's main contribution is on the measurement side. The operating assumption that guides the empirical work is that the managerial forecast errors are (inversely) proportional to the stock of data available to firms. This assumption is predicated on the following: 1) firms that are more likely to use data in production have lower forecast errors; 2) managerial forecasts carry information that is additive to the information of other market participants. Here, I will discuss both of these assumptions in order.

How are firms using data in production?

In the model, firms use data to directly increase profits. In particular, the firm's flow profits depend on the deviation between a specific action and a state which is imperfectly observed. Better data allows the firm to take the action that is most appropriate to the aggregate state, thereby increasing profits. As an example, the firm's action is the choice of how much to produce in a given state of the world, and the state to be forecast is a function of demand for the firm's product. Minimizing the distance between how much the firm produces and demand maximizes profits. The assumption that maps the model to the data is that management forecast errors of future revenue are a good proxy for the firm's forecast errors. One can ask whether the model meaningfully captures the dimensions along which firms use data to increase profits, and also the extent to which managerial forecast errors are correlated with the underlying stock of data available to firms.

In practice, firms use data in production in a variety of ways. Table 1 summarizes a few examples from firms' regulatory filings. The *Stitch Fix* example is perhaps closest to the heart of the model: the firm is using data provided by customers to provide more targeted recommendations and also forecast demand, design new products and manage inventory. *John Deere* and *Kroger* use customer data to forecast inventory needs and provide more targeted recommendations, respectively. Additionally, *Kroger* sells customer data to third parties, who presumably use it for serving targeted advertising. *Affirm*, a financial firm, uses customer data to more effectively price credit risk for applicants, which at the same time also allows them to price discriminate—increasing their monopoly power and 'creating a moat' around the business. Last, *Klarna* uses data to train AI customer service agents which allows them to save on labor costs. These examples illustrate that firms use data partly to

forecast customer demand, but they also use data in ways that are somewhat outside the model, specifically enhance their monopoly power and also develop labor-saving technologies. It is not entirely obvious that the degree to which firms do the latter is positively correlated with the quality of their managerial forecast errors.

As such, I think the paper could benefit from a broader validation exercise. Currently, the authors show in Section 5 that managerial forecast errors are (inversely) related to firm employment of data scientists. Though useful, there are alternative interpretations of this correlation that do not rely on data: perhaps better managers make more precise forecasts, and at the same time, these managers are investing in employees with a data science background. Further, focusing on data scientists is somewhat inconsistent with the paper's motivation, since the wages of data scientists do appear in national accounts. What the paper is really about is firms acquiring data as a by-product of their operating decisions, specifically receiving data from customers in exchange for an (implicit) price discount.

My suggestion for the authors is to use modern NLP techniques on firms' 10K documents—which are publicly available starting from 1993—and verify the extent to which firms with lower managerial forecast errors relative to their peers are also more likely to use customer data in forecasting consumer demand. Showing a positive correlation would validate the central implication of the model: managerial forecast errors are an indirect proxy for the firm's stock of data.

Managerial forecast errors can be biased

The mapping from forecast errors to the data stock presumes that managers form rational, Bayesian forecasts—that reported guidance equals the conditional expectation of revenue, so that the squared forecast error measures the conditional variance the firm faces and hence its stock of data. A growing body of work on managerial expectations casts doubt on this premise. Ma, Ropele, Sraer, and Thesmar (2024) study a long panel of sales forecasts from a representative sample of Italian firms and find that, while forecasts are unconditionally unbiased, they are conditionally biased: forecast errors are predictable from their own lag, a pattern the authors interpret as managers' under-reaction to information. Bordalo, Gennaioli, Shleifer, and Terry (2026) reach a related conclusion from the opposite direction: disciplining beliefs with US managers' expectations about their own firms' profitability, they find that expectations over-react to recent news. The two papers disagree on the sign of the distortion, but they agree on the substantive point: managerial forecast errors are not a clean reflection

of the information available to the firm, but embed a systematic, predictable wedge between the rational forecast and the one managers report.

This matters for the measurement exercise because the realized squared forecast error confounds two distinct objects. The mean squared forecast error decomposes into the conditional variance of the state—the object the paper seeks, governed by the firm's stock of data—and the variance of the forecast bias, which reflects how managers process information rather than how much of it they hold. A firm whose managers overreact to recent news, or whose errors are predictable from past mistakes, will appear data-poor in the paper's metric even when its information set is rich. Because the distortions these papers document vary across firms and over time—and are plausibly largest during turbulent periods such as 2008 and 2020—the inferred data stock will inherit this variation. A natural robustness check is to extend the empirical procedure to allow for the possibility of biased forecasts. For instance, one could purge the predictable component of forecast errors before mapping them into precision: project errors on the variables that proxy for the distortion—their own lags, recent profit or news surprises, or a calibrated diagnosticity parameter in the spirit of Bordalo et al. (2026)—and use the residual, unpredictable error as the measure of the conditional variance. Such a correction would separate the part of a forecast miss that the firm could not have avoided given its data from the part that reflects how its managers weight that data, lending the resulting data-stock series a cleaner interpretation.

2.2 Identification

The value the paper assigns to data is disciplined largely by the estimated relation between gross profits and forecast errors in Table 3, which the calibration interprets causally: a reduction in forecast errors, achieved by acquiring more data, raises gross profits, and the slope of that relationship maps into γ, the marginal value of data. The identification challenge is that the same slope is consistent with alternative models in which data plays no role: a superior management team produces more accurate guidance and earns higher profits because it makes better decisions across many margins—pricing, hiring, investment, and product design—of which forecasting is only one. This omitted dimension of firm quality generates a negative correlation between forecast errors and profits that does not operate through the data channel, biasing the estimated slope upward in magnitude and, with it, the imputed value of data and its overall contribution to GDP. A related channel is scope: larger and more profitable firms operate more diversified product lines whose aggregate revenue is mechanically more predictable, again producing a profit–forecast-error relationship that is not about data.

Both of these concerns are distinct from the endogeneity problem that the paper's recursive framework is designed to address. The model deals with a specific channel of reverse causality: higher profits at t allow a firm to generate more data, which lowers its forecast errors and raises profits again at $t+1$. This dynamic feedback is disciplined by the second regression, of forecast errors on lagged profits in Table 4, and is precisely what the calibration is built to account for when it matches the slopes of both regressions across simulated and real data.

Is this a first order concern? In response to my comments, the authors have added firm fixed effects to the last column of Table 3. This is a meaningful step, as it absorbs the permanent component of management quality; doing so reduces the estimated coefficient by approximately 50%, though the estimate is still highly statistically significant. However, firm fixed effects only partly alleviate this concern: the within-firm variation that now identifies the slope still mixes the data channel with time-varying quality—the arrival of a new chief executive, a reorganization, or a demand boom that is at once more predictable and more profitable—as well as with within-firm changes in scope.

Mitigating this issue requires quasi-exogenous variation in either the quality of data or the technology available for processing data. The identification assumption would be that the variation in a firm's data, or in the precision of its forecasts, is plausibly orthogonal to its management quality. One possibility is an instrument for data-related hiring based on pre-existing hiring networks (Hampole, Papanikolaou, Schmidt, and Seegmiller, 2025). Specifically, Hampole et al. (2025) instrument firms' exposure to data- and AI-related work using pre-existing hiring networks—historical university-to-firm pipelines that shift a firm's access to technical talent for reasons unrelated to its current decisions. The same design could be applied here: instrument a firm's hiring of data workers, and hence the precision of its forecasts, with the supply of data-skilled graduates flowing through its historical hiring channels. The first stage would link the instrument to forecast accuracy, and the second would recover the causal effect of precision on profits—a direct estimate of the object the calibration needs, rather than one read off a potentially confounded slope.

A second possibility is changes in regulation. Changes in data and privacy law shift the availability and usability of customer data differentially across firms and over time: the European Union's General Data Protection Regulation, or state-level privacy statutes such as the California Consumer Privacy Act, curtail the data that exposed firms can collect or retain—by sector, by reliance on third-party data, or by the geography of their customers—in a way that is plausibly independent of any one firm's management quality. A difference-in-differences design comparing more- and less-exposed firms around these reforms would

identify how a contraction in data access raises forecast errors and depresses profits, providing an out-of-sample check on the central elasticity of the model.

2.3 Non-Gaussian uncertainty and tail risk

The tractability of the model rests on the joint assumption of a quadratic loss function together with a Gaussian state and noise. These assumptions imply that firms' expected profit depends on the forecast only through its conditional variance. Hence, the stock of data is summarized by precision and the value of data by the reduction in variance, with the squared forecast error as the empirical counterpart. Though tractable, this structure imposes some strong restrictions: it implies that firms are indifferent between two information sets that deliver the same forecast-error variance but different tail behavior. In practice firms care disproportionately about avoiding large misses—stockouts, supply-chain breaks, liquidity shortfalls—whose cost is convex and often asymmetric, in that a shortfall hurts more than an equivalent overshoot.

To see this point more precisely, consider a simple extension of the firm's decision problem in the paper. A firm chooses an action a before a payoff-relevant state θ is realized. Conditional on its information set $\mathcal{I}$, it holds beliefs over θ with mean $\mu = \mathrm{E}[\theta \mid \mathcal{I}]$, variance σ^2, and higher central moments $\mu_k = \mathrm{E}[(\theta - \mu)^k \mid \mathcal{I}]$. Flow profit is

$$\pi(a, \theta) = \bar{\pi} - L(a - \theta), \tag{1}$$

where the loss L is smooth, with $L(0) = 0$ and $L'' > 0$; the paper is the special case of a quadratic loss, $L(x) = \frac{\gamma}{2}x^2$. The per-period value of data $\mathcal{I}$ is therefore given by,

$$V(\mathcal{I}) = \bar{\pi} - \min_a \ \mathrm{E}[L(a - \theta) \mid \mathcal{I}]\,, \tag{2}$$

and the value of new data is $V(\mathcal{I}') - V(\mathcal{I})$ when the firm refines its information $\mathcal{I}$ to $\mathcal{I}'$. Letting $\delta = a - \mu$ denote the action relative to the forecast and expanding the loss in the forecast error gives

$$\mathrm{E}[L(a - \theta) \mid \mathcal{I}] = L(\delta) + \tfrac{1}{2}L''(\delta)\,\sigma^2 - \tfrac{1}{6}L'''(\delta)\,\mu_3 + \tfrac{1}{24}L''''(\delta)\,\mu_4 + \cdots \tag{3}$$

This general setting nests the model in the paper: there are two cases under which the conditional variance is a sufficient statistic for the value of data. First, under a quadratic loss, the firm simply sets $\delta^* = 0$ and $V = \bar{\pi} - \frac{\gamma}{2}\sigma^2$, as in the paper, regardless of the data-generating process. Second, under normality, the forecast error is normal with variance σ^2 for

any action, so V depends on beliefs only through σ^2 for *any* loss L; asymmetry then changes the optimal action and the level of profit, but the value of data is still determined by the conditional variance.

Consider, however, the case in which the loss function is asymmetric and the data-generating process is non-normal. In this case, the stock of data can no longer be inferred from the conditional variance of the firm's forecast errors. To see this, consider the third term in (3). A firm whose loss function exhibits $L''' < 0$ (a shortfall costs more than an equal overshoot) is averse to left-skewed forecast errors and values data that thins their lower tail—lowering the chance of a large shortfall—holding fixed the variance of the forecast error. The next term ties the fourth derivative of the loss function (L'''') to the kurtosis μ_4 of the forecast errors, and so on. The sufficient statistic for the stock of data is not the squared forecast error but the full set of conditional moments $(\sigma^2, \mu_3, \mu_4, \dots)$, weighted by the curvature of profit at the optimum. Estimating the stock of data is now considerably more involved: it requires estimating additional parameters and uses the full distribution of forecast errors, and hence should be outside the scope of the current paper.

Would this modification lead to meaningfully different results? I think so, and the authors' own results point to this direction. As we see in Figure 1 in the paper, the distribution of squared forecast errors is heavily right-skewed. The authors accommodate the skew through rescaling the estimates so that they correspond to the median firm. If the value of data lies in that right tail, however—if data is valuable precisely because it occasionally averts a disaster—then this rescaling may be discarding the most informative variation in the data. We can also see this in Figure 4, which plots the estimated value of data per employee for the median firm in the sample. In the figure, 2020 reads as if data has a negative value added; this is driven by the fact that realized squared errors spiked and the flow of new data fell. An alternative interpretation of the same facts, however, can lead to different conclusions: 2020 is the period when resolving downside uncertainty was most valuable, and the spike in squared errors reflects a fat-tailed shock rather than a reduction in the value of data.

3 Summary

Measuring the contribution of intangible assets to economic output is a first order question, and a large component of intangibles is data. The paper isolates a component of data's value that the national accounts miss twice—once as intangible investment, once as barter—and proposes a novel way of recovering it based on the quality of firms' revenue forecasts. My comments

focus on clarifying the conditions under which the authors' measure of the stock of data can be taken at face value and should be read as suggestions for future work on this important agenda.

References

Simona Abis and Laura Veldkamp. The changing economics of knowledge production. *The Review of Financial Studies*, 37(1):89–118, 2024.

Pedro Bordalo, Nicola Gennaioli, Andrei Shleifer, and Stephen J. Terry. Real credit cycles. *American Economic Review*, 116(4):1274–1308, April 2026. doi: 10.1257/aer.20211820.

Maryam Farboodi and Laura Veldkamp. A model of the data economy. Working Paper 28427, National Bureau of Economic Research, 2021.

Menaka Hampole, Dimitris Papanikolaou, Lawrence D. W. Schmidt, and Bryan Seegmiller. Artificial intelligence and the labor market. Working Paper 33509, National Bureau of Economic Research, 2025.

Charles I. Jones and Christopher Tonetti. Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–2858, 2020.

Yueran Ma, Tiziano Ropele, David Sraer, and David Thesmar. A quantitative analysis of distortions in managerial forecasts. Working paper, 2024.

Tables and Figures

Table 1: Specific examples of firms' data use in operations

Stitch Fix (SFIX)	2024 10K	Stitch Fix is powered by data science. Our data science capabilities consist of our rich data set, best-in-class artificial intelligence and our proprietary algorithms, which fuel our business by enhancing the client experience and driving business model efficiencies. The vast majority of our client data is provided directly and explicitly by the client, rather than inferred, scraped, or obtained from other sources. We also gather extensive merchandise data, such as inseam, pocket shape, silhouette, and fit. This large and growing data set provides the foundation for proprietary algorithms that we use throughout our business, including those that make recommendations to our Stylists, predict purchase behavior, forecast demand, optimize inventory, and enable us to design new apparel.
John Deere (DE)	2025 10K	Advancements such as precise global navigation satellite systems technology, advanced connectivity and telematics, on-board sensors and computing power, automation software, digital tools, applications, and analytics provide seamless integration of information designed to improve customer decision-making and job execution. To ensure adequate inventory supply, we must forecast inventory needs and expenses and place orders sufficiently in advance with suppliers and contract manufacturers.
Kroger (KR)	2024 10K	The traffic and data generated by our retail business, including pharmacies and fuel centers, enables our diverse business. Kroger serves approximately 63 million households annually and because of our rewards program, over 95% of customer transactions are tethered to a Kroger loyalty card. Our over 20 years of investment in data science capabilities allows us to utilize this data to create personalized experiences and value for our customers and enables our growing, high operating margin alternative profit businesses, including data analytic services and third-party media revenue.

Affirm Holdings (AFRM)	2025 10K	Our risk model takes five top-of-mind data inputs from the user and turns them into a total of over 500 data points in order to assess the credit risk of new consumers. Our algorithms model out the repayment probability on a month-to-month basis, and combine these probabilities with the term length, purchase size, merchant, and item being purchased, in order to price and score risk. In the vast majority of cases, we can complete these checks and calculations in a matter of seconds... We believe our technology, underwriting, and risk management are key competitive advantages. Our proprietary technology's ability to price and assess risk at a transaction level provides a unique advantage compared to legacy payment and credit systems. Furthermore, our risk management models are designed to continuously improve over time, becoming more precise and efficient with each transaction. This translates into increased purchasing power with more control and flexibility for consumers. By utilizing our unique risk model predicated on sophisticated machine learning algorithms, proprietary data, and product-level underwriting, we can serve consumers across the credit spectrum and price risk across transaction types ... From day one, Affirm was designed to build towards network effects of data. Routing information-rich messages allows us to rapidly improve the products we offer, ensuring the value curve keeps rising and creating a moat around our business. The rich data we gather during each transaction's lifetime improves our many models: credit and fraud, AdaptAI (personalized incentives to maximize sales conversion), BoostAI (program selection for merchants), etc. As we refine these models using data other lenders can't access, we improve credit approvals and consumer take-up at the point of sale.
Uber (UBER)	2025 10K	We have built proprietary marketplace, routing, and payments technologies. Marketplace technologies are the core of our deep technology advantage and include demand prediction, matching and dispatching, and pricing technologies. Our technologies make it extremely efficient to launch new businesses and operationalize existing ones. Our revenue is dependent on the pricing models we use to calculate consumer prices and Driver earnings. Our pricing models, including dynamic pricing, have been, and will likely continue to be, challenged, banned, limited in emergencies, and capped in certain jurisdictions. For example, we have agreed to not calculate consumer fares in excess of the maximum government-mandated fares in all major Indian cities where legal proceedings have limited the use of surge pricing.

Klarna	F-1 Registration Statement (March 2025, 333-285826)	Our network and AI capabilities are powered by a unique proprietary data set, built on SKU-level data points, including over 2.5 billion data points collected in 2024, and more than 5.2 billion transactions conducted through our network since our founding. Our AI assistant is handling 62% of customer service chats, according to our service chat log data, doing the work equivalent of over 800 full-time agents (estimated based on the average monthly reduction in chat and telephone conversations handled by full-time agents following the launch of our AI assistant), and in 2024 delivered approximately $39 million in cost savings.